Экзаменационный проект по КИЛИ

Corpus of Historical American English https://corpus.byu.edu/coha/

Афлатунова Александра(aiaflatunova@gmail.com),

Рачинская Софья (sophya.rach@gmail.com)

Введение

Для экспертизы мы выбрали Исторический Корпус Американского Английского (COHA). В нем содержатся тексты на американском английском с 1810 года до настоящего времени. Это самый большой структурированный исторический корпус английского. В нем содержатся более чем 400 миллионов слов в 100 тысячах различных текстов. Корпус создан по гранту от National Endowment for the Humanities.

Usability

Дизайн

Сайт ресурса оформлен просто и практично. Цвета приятны глазу, но при этом не очень продумано их использование - вся представленная информация выглядит однообразно. В остальном дизайн довольно практичен и сосредоточен на функциональности. Имеется версия для мобильных устройств не сильно отличающаяся от основного сайта. В ней обнаруживается пара недочетов как, например, невозможность свободного масштабирования, из-за чего неудобно работать с контекстом. Несмотря на все инфографика представлена достаточно наглядно и удобна в навигации.

Onboarding

Для новичка сайт изначально может показаться простым в использовании в силу простоты дизайна. На самом деле, без прочтения руководства по использованию ресурса в нем сложно разобраться интуитивно. Навигация довольно запутана т.к. предполагается, что пользователь будет использовать перемещение по каталогам вместо привычных *обратно* и *вперед* браузеров. Это частично восполняется тем, что окно помощи всегда присутствует рядом с окном поиска и показывает актуальную справку по любой выбранной операции, включая примеры запросов.
Начать что-то искать очень легко, так как окно находится на главной странице. Разобраться же помогает окно помощи, в котором также можно найти ссылку на гайд по использованию ресурса (на который также можно выйти через вкладку TOUR вверху страницы). К сожалению эти окна никак не отличаются от остальных и все ссылки выглядят абсолютно одинаково, так что пользователь должен начать читать справку чтобы их обнаружить. Так что для полноценного пользования ресурсом нужно сначала найти и внимательно прочитать инструкцию.

Функционал

- *Доступность*

  Корпус доступен онлайн, но с ограничением запросов до 15 в день без регистрации и 50 для студентов с регистрацией. Большее количество запросов доступно для ученых, преподавателей или для подписчиков.
  Есть возможность скачать корпус целиком (платная).

- *Слои разметки:*

  Тексты включают в себя как морфологическую так и метаразметку (указана вся информация о текстах, например, жанр, страна, источник, автор, год написания).
  Нет синтаксической разметки, а из семантики есть только синонимия.

- *Состав*

  Корпус сбалансирован по жанрам на каждый период времени (например, художественная литература составляет 48-55% из всех текстов за одно десятилетие
  Содержит 107000 текстов написанных в период с 1810 года до настоящего момента

- *Тексты, включенные в корпус:*

  Fiction: Project Gutenberg (1810-1930), Making of America (1810-1900), сканы книг (1930-1990), сценарии фильмов и театральных постановок, COCA (1990-2010)
  Журналы: Making of America (1810-1900), сканы и PDF (1900-1990), COCA (1990-2010)
  Газеты: PDF > TXT из исторических газетных архивов (1850-1980), COCA etc (1990-2010)
  Non-fiction: Project Gutenberg (1810-1900), www.archive.org (1810-1900), сканы книг (1900-1990), COCA (1990-2010)

- *Возможности*

  Ресурс позволяет производить поиск по словам, фразам, леммам. Кроме этого возможен поиск по началу и концу слова (примерно как в НКРЯ) а также более сложные запросы включающие поиск фраз определенной конструкции (типа «слово» + прил + сущ)
  В качестве результата можно вывести частотность появления запрашиваемой конструкции в текстах из года в год (при этом контекст будет доступен кликом мышки по слову), также можно вывести слова вместе с которыми встречается каждая запрашиваемая словоформа при сравнении двух.
  Возможны сравнение использования слов по временным периодам
  Можно искать слово с его синонимами
  Можно использовать пользовательский список слов, что частично решает проблему отсутствия семантической разметки
  Кроме того, можно задавать подкорпусы по метапризнакам текста, например, жанра, года создания и пр.
  К сожалению мы не нашли возможность скачать выборку, но при наличии аккаунта результаты можно сохранять себе в профиль (количество сохраняемых результатов зависит от уровня аккаунта).
  Также в той же системе находятся ещё несколько корпусов и есть возможность сравнивать результаты запросов в них, что тоже может быть очень удобно (к примеру может пригодиться возможность сравнить результаты в корпусах исторического и современного американского английского).

Примеры работы с корпусом

Попробуем, например, посмотреть, какие слова чаще всего стоят перед словосочетанием *sliced bread*, для этого используем вот такой запрос:

Полученный результат:

		CONTEXT	FREQ		ALL
1		SINCE	8		183516
2		THING	8		188215
3		BEST	5		147501
4		LOAF	4		2528
5		THINLY	2		1481
6		GREATEST	2		33851
7		KNAL	1		1
8		NEATEST	1		228
9		$|P75	1		368
10		SALADS	1		611
11		GROUSE	1		934
12		LARD	1		989
13		CHUNKS	1		1261
14		TRAYS	1		1275
15		PLATTER	1		1354
16		PEACH	1		2399

Заметно, что чаще всего *sliced bread* используется как часть идиомы *best/greatest thing since sliced bread*.

Теперь мы можем посмотреть на то, в каких контекстах использовалось сочетание *since sliced bread*:

1	1962	MAG	Time	A	B	C	floor clean and dry. Says Family Circle Magazine: " The greatest thing since sliced bread. " Price: $
2	1980	MAG	NewYorker	A	B	C	a promotional leaflet we picked up at the symposium? " the greatest thing since sliced bread. " W
3	1981	NEWS	Boston	A	B	C	COMING!!!!? L hought it was the best thing since sliced bread, " Jim Baker recalled. The voters. In
4	1988	FIC	Mov:RunningOnEmpty	A	B	C	LORNA (against this instruction) My father thinks you're the best thing since sliced bread. DANN
5	1995	FIC	FantasySciFi	A	B	C	with sickeningly sweet fruit fillings; I thought red pepper was the neatest thing since sliced bread.
6	1997	NEWS	NYT	A	B	C	their best side. To me back then, Texaco was the best thing since sliced bread. " # Texaco also off
7	1999	MAG	ConsumRep	A	B	C	think it causes cancer and other people who think it's the best thing since sliced bread. " # To pro
8	2006	NEWS	Chicago	A	B	C	People for the Ethical Treatment of Animals, calls it " the best thing since sliced bread. " Friedrich,

Увидим, что впервые эта идиома видимо появилась в шестидесятые как *greatest thing since sliced bread*, а в восьмидесятые приобрела популярность, при этом *greatest* стали заменять на *best*.

Другая возможность, которую предоставляет этот корпус - это сравнение слов по их контекстам, к примеру, посмотрим какие существительные встречаются после слов *modern* и *contemporary* и отсортируем по частоте использования:

WORD 1 (W1): **CONTEMPORARY** (0.18)

	WORD	W1	W2	W1/W2	SCORE
1	[ART]	336	1118	0.3	1.7
2	[LIFE]	119	764	0.2	0.9
3	[WRITER]	109	202	0.5	3.0
4	[SOCIETY]	91	660	0.1	0.8
5	[HISTORY]	83	574	0.1	0.8
6	[ARTIST]	81	99	0.8	4.5
7	[WORLD]	80	1051	0.1	0.4
8	[MUSIC]	68	141	0.5	2.7
9	[WORK]	65	115	0.6	3.1
10	[LITERATURE]	64	225	0.3	1.6
11	[REVIEW]	46	0	92.0	509.3
12	[CULTURE]	40	115	0.3	1.9
13	[ACCOUNT]	39	4	9.8	54.0
14	[NOVELIST]	38	46	0.8	4.6
15	[CRAFT]	36	5	7.2	39.9
16	[HISTORIAN]	36	90	0.4	2.2
17	[EVENT]	35	4	8.8	48.4
18	[PROBLEM]	35	39	0.9	5.0
19	[RECORD]	34	16	2.1	11.8
20	[CRITICISM]	34	43	0.8	4.4
21	[FICTION]	34	86	0.4	2.2
22	[ISSUE]	33	4	8.3	45.7
23	[THINK]	33	176	0.2	1.0
24	[POET]	30	108	0.3	1.5
25	[PAINTING]	30	139	0.2	1.2
26	[DOCUMENT]	28	3	9.3	51.7
27	[CIVILIZATION]	28	513	0.1	0.3
28	[SCIENCE]	28	675	0.0	0.2
29	[SCENE]	27	10	2.7	14.9
30	[OPINION]	27	14	1.9	10.7

WORD 2 (W2): **MODERN** (5.54)

	WORD	W2	W1	W2/W1	SCORE
1	[TIME]	2001	11	181.9	32.9
2	[ART]	1118	336	3.3	0.6
3	[WORLD]	1051	80	13.1	2.4
4	[LIFE]	764	119	6.4	1.2
5	[SCIENCE]	675	28	24.1	4.4
6	[SOCIETY]	660	91	7.3	1.3
7	[HISTORY]	574	83	6.9	1.2
8	[CIVILIZATION]	513	28	18.3	3.3
9	[LANGUAGE]	459	4	114.8	20.7
10	[MAN]	455	22	20.7	3.7
11	[STATE]	250	12	20.8	3.8
12	[CITY]	241	3	80.3	14.5
13	[DAY]	231	0	462.0	83.5
14	[LITERATURE]	225	64	3.5	0.6
15	[METHOD]	220	4	55.0	9.9
16	[WAR]	203	4	50.8	9.2
17	[WRITER]	202	109	1.9	0.3
18	[WARFARE]	199	1	199.0	35.9
19	[AGE]	198	4	49.5	8.9
20	[THINK]	176	33	5.3	1.0
21	[IDEA]	175	13	13.5	2.4
22	[IMPROVEMENT]	173	0	346.0	62.5
23	[TECHNOLOGY]	168	7	24.0	4.3
24	[ARCHITECTURE]	168	15	11.2	2.0
25	[MEDICINE]	167	5	33.4	6.0
26	[SCHOOL]	167	8	20.9	3.8
27	[NATION]	160	7	22.9	4.1
28	[BUILDING]	159	8	19.9	3.6
29	[INDUSTRY]	157	1	157.0	28.4
30	[SENSE]	153	15	10.2	1.8

Заметно, что, например, *life*, *society* и слова, связанные с искусством, часто употребляются и с тем, и с другим прилагательным. С другой стороны, в связи с документами чаще можно встретить слово *contemporary*. Это же видно, если отсортировать выдачу по соотношению:

WORD 1 (W1): **CONTEMPORARY** (0.18)

	WORD	W1	W2	W1/W2	SCORE
1	REVIEW	45	0	90.0	498.2
2	EVIDENCE	22	0	44.0	243.6
3	ACCOUNT	14	0	28.0	155.0
4	DOCUMENTS	20	1	20.0	110.7
5	FILMS	10	0	20.0	110.7
6	CHEFS	10	0	20.0	110.7
7	CRISIS	7	0	14.0	77.5
8	REPUTATION	7	0	14.0	77.5
9	WITNESSES	7	0	14.0	77.5
10	CRAFTS	24	2	12.0	66.4
11	CHRONICLERS	6	0	12.0	66.4
12	CHRONICLER	6	0	12.0	66.4
13	EVENTS	34	3	11.3	62.7
14	TESTIMONY	10	1	10.0	55.4
15	MEMOIRS	5	0	10.0	55.4
16	COST	5	0	10.0	55.4

WORD 2 (W2): **MODERN** (5.54)

	WORD	W2	W1	W2/W1	SCORE
1	LANGUAGES	327	0	654.0	118.1
2	DAYS	181	0	362.0	65.4
3	IMPROVEMENTS	143	0	286.0	51.7
4	TIMES	1926	9	214.0	38.7
5	WARFARE	199	1	199.0	35.9
6	EQUIPMENT	81	0	162.0	29.3
7	CONVENIENCES	73	0	146.0	26.4
8	INDUSTRY	144	1	144.0	26.0
9	HUMANS	69	0	138.0	24.9
10	INVENTION	64	0	128.0	23.1
11	NATION	63	0	126.0	22.8
12	WEAPONS	120	1	120.0	21.7
13	MACHINERY	59	0	118.0	21.3
14	DAY	50	0	100.0	18.1
15	ECONOMY	49	0	98.0	17.7
16	CITY	194	2	97.0	17.5

По получившимся результатам (и их контексту) можно понять, что хотя оба слова переводятся на русский язык как *современный*, у них немного разный смысл: *contemporary* означает *современный чему-то*, *из одной эпохи*, в то время как *modern* - *современный говорящему*, *нынешний*.

Вывод

Несмотря на то, что работать с корпусом немного сложнее чем, например, с НКРЯ, у него есть много удобных и полезных функций, которые могут помочь как в исследованиях так и просто при изучении английского языка (например, когда требуется выбрать из двух синонимов более подходящий).
К сожалению, отсутствие возможности скачать результаты запроса могут стать препятствием в исследованиях, увы, интерфейс корпуса не может дать всех возможностей, например, Excel-я.